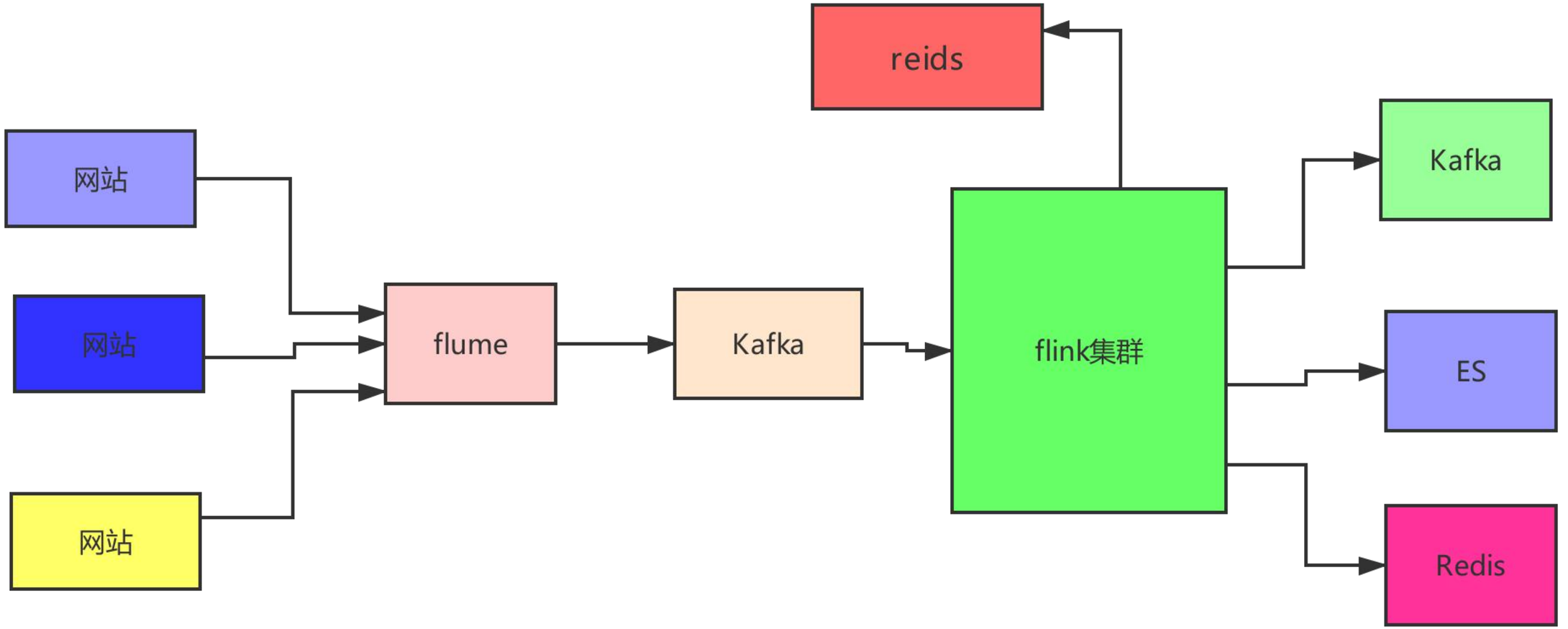
二手电商用户行为分析（架构）

01 项目流程



01 实时热门页面统计

需求分析

对日志数据进行实时ETL

数据展示

dt	countryCode	type	score	level
24.233.162.179	IN	s5	0.2	D

01 效果展示

原始数据格式


```
{"dt":"2019-11-19 20:33:39","countryCode":"TW","data":[{"type":"s1","score":0.8,"level":"D"},
{"type":"s2","score":0.1,"level":"B"}]}

{"dt":"2019-11-19 20:33:41","countryCode":"KW","data":[{"type":"s2","score":0.2,"level":"A"},
{"type":"s1","score":0.2,"level":"D"}]}

{"dt":"2019-11-19 20:33:43","countryCode":"HK","data":[{"type":"s5","score":0.5,"level":"C"},
{"type":"s2","score":0.8,"level":"B"}]}
```


效果数据格式


```
{"dt":"2019-11-19 20:33:39","area":" AREA_CT", "type":"s1","score":0.8,"level":"D"}
{"dt":"2019-11-19 20:33:39","area":" AREA_CT","type":"s2","score":0.1,"level":"B"}
```


码表数据（redis,大区）

```
hset areas AREA_US US
hset areas AREA_CT TW,HK
hset areas AREA_AR PK,KW,SA
hset areas AREA_IN IN
```

奈学教育

01 项目背景

02 项目架构/架构选型

03 需求实现

第一个需求分析

01 项目背景

一个二手交易平台

一个帮你赚钱的网站

每年帮助超过1000万用户卖出宝贝


中国联通 4G
下午 5:17
转转二手交易网
华为平板
热搜：
限时加价10%
airpods
苹果6
iphonex
小米
双11Super福利提前抢
预付1元可抵388·爆款手机免费抽
专业质检
已完成68781272次验机服务
专业智能验机
放心严选好货
手机3C保卖
腾讯战略投资
苹果手机
安卓手机
平板
电脑
特价零食
文玩玉翠
自营图书
二手车
限时秒杀 15:42:22
全部秒杀
办公无线鼠标
首页
省心卖
购物车
我的


中国联通 4G
下午 5:31
转转二手交易网
华为平板
热搜：
限时加价10%
小米9
iphonex
airpods
小米
老罗推荐
转转严选二手手机
官方验机 | 1年质保 | 7天无理由退货
专业质检
已完成68781845次验机服务
专业智能验机
放心严选好货
手机3C保卖
腾讯战略投资
苹果手机
安卓手机
平板
电脑
特价零食
文玩玉翠
自营图书
二手车
限时秒杀 15:28:17
全部秒杀
办公无线鼠标
首页
省心卖
购物车
我的


中国联通 4G
下午 5:30
转转自营二手书
道德经
¥9.00
原价 36
2.5折
全新书
新人
新人30元大礼包
满49减12元
满29减5元
领取
自营图书
道德经
1人卖过这本书
服务
自营图书
消毒塑封
100%正版
满45包邮
作者
(春秋)老子
更多作品
出版社
华龄出版社 2017-06
更多作品
首页
购物车
加入购物车
立即购买

1 点击/浏览

2 加入购物车

3 下单

4 支付

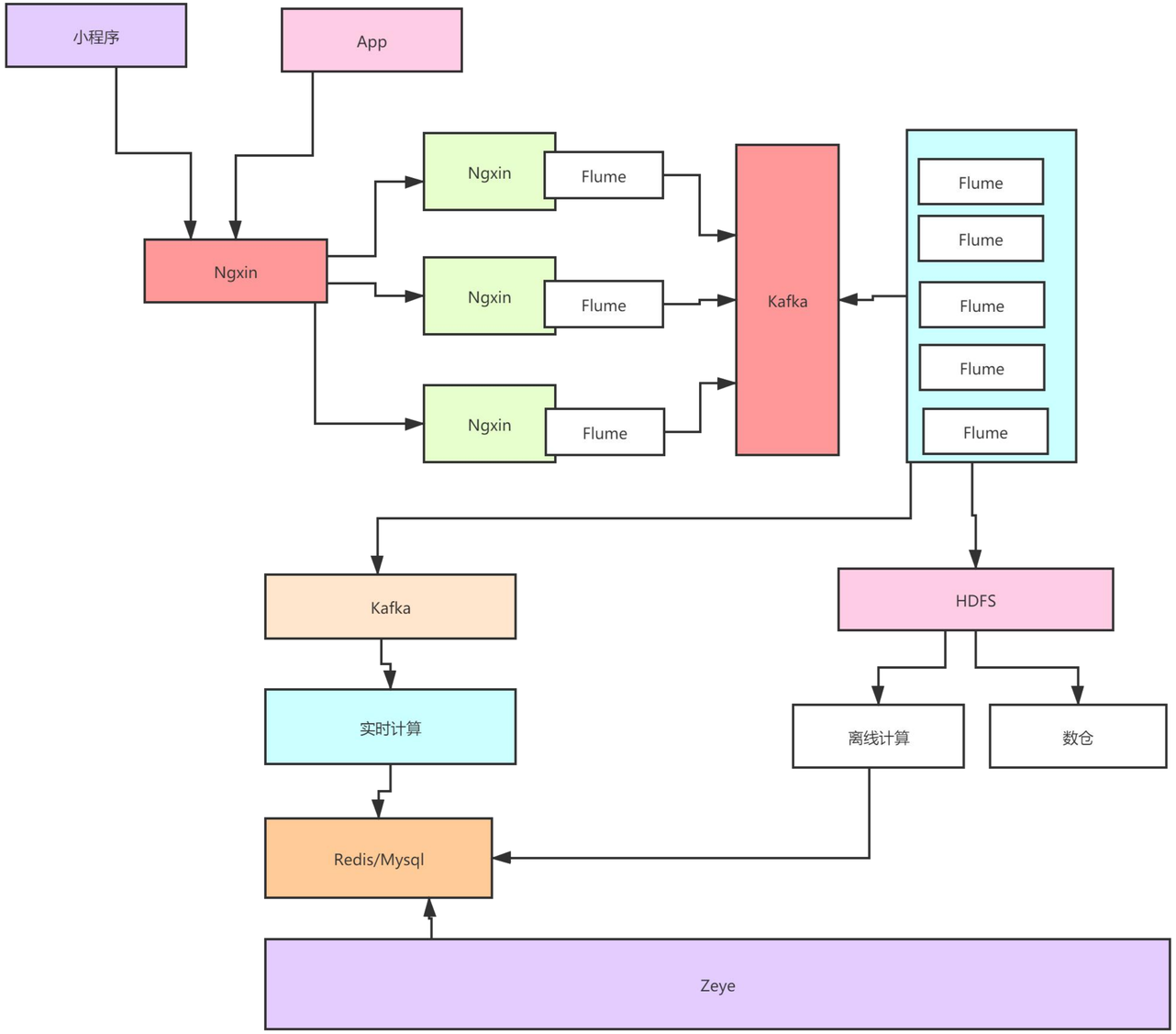
5 评价

Zeye系统

转转属于电商平台，电商平台的用户行为频繁且较复杂，系统上线后每天收集到大量的用户行为数据， 公司需要一个系统利用大数据技术进行深入挖掘和分析， 得到感兴趣的商业指标用来做数据分析和商业决策，并增强对风险的控制，故转转开发了Zeye系统。

02 项目架构/架构选型

02 项目架构/技术选型


小程序
App
Ngxin
Ngxin
Flume
Ngxin
Flume
Ngxin
Flume
Kafka
Flume
Flume
Flume
Flume
Flume
Kafka
HDFS
实时计算
离线计算
数仓
Redis/Mysql
Zeye

03

热门页面统计

01 实时热门页面统计

需求分析

每隔5秒统计最近10分钟热门页面

数据展示

IP地址	用户ID	事件时间	请求方式	URL
24.233.162.179	-	17/05/2020:11:05:31 +0000	GET	/images/jordan-80.png

实现思路

01 效果展示

```
时间: 2020-05-20 21:13:25.0
URL:/blog/tags/puppet?flav=rss20 访问量: 6
URL:/projects/xdotool/ 访问量: 4
URL:/favicon.ico 访问量: 4
URL:/images/web/2009/banner.png 访问量: 3
URL:/presentations/logstash-puppetconf-2012/css/reset.css 访问量: 3
====================
时间: 2020-05-20 21:13:30.0
URL:/blog/tags/puppet?flav=rss20 访问量: 6
URL:/projects/xdotool/ 访问量: 4
URL:/favicon.ico 访问量: 4
URL:/images/web/2009/banner.png 访问量: 3
URL:/presentations/logstash-puppetconf-2012/css/reset.css 访问量: 3
====================
```

01 思考?

需求分析

每隔5秒统计最近10分钟热门品类/商品?

01 实时统计热门商品

需求分析

每隔5分钟统计最近1个小时热门商品

数据展示

用户编号	商品编号	品类编号	用户行为	访问时间	会话id
4675130	12437439	4296588	P	1601688552	dfac5866-f644-4cb5-a303-e941fb331f77

注：原来的数据格式是json格式


1
读取数据
2
添加水位
3
过滤用户行为
4
按商品分组
5
统计窗口数据
6
根据窗口分组
7
商品TopN排序
8
打印输出

实现思路

03 效果展示

```
时间: 2020-10-03 09:40:00.0
 商品ID: 21123024 商品浏览量=6
 商品ID: 14188074 商品浏览量=5
 商品ID: 3104010 商品浏览量=5
==============================
时间: 2020-10-03 09:45:00.0
 商品ID: 21123024 商品浏览量=8
 商品ID: 1219464 商品浏览量=7
 商品ID: 3104010 商品浏览量=6
==============================
时间: 2020-10-03 09:50:00.0
 商品ID: 1219464 商品浏览量=9
 商品ID: 20837616 商品浏览量=8
 商品ID: 21123024 商品浏览量=8
```

04 广告点击统计

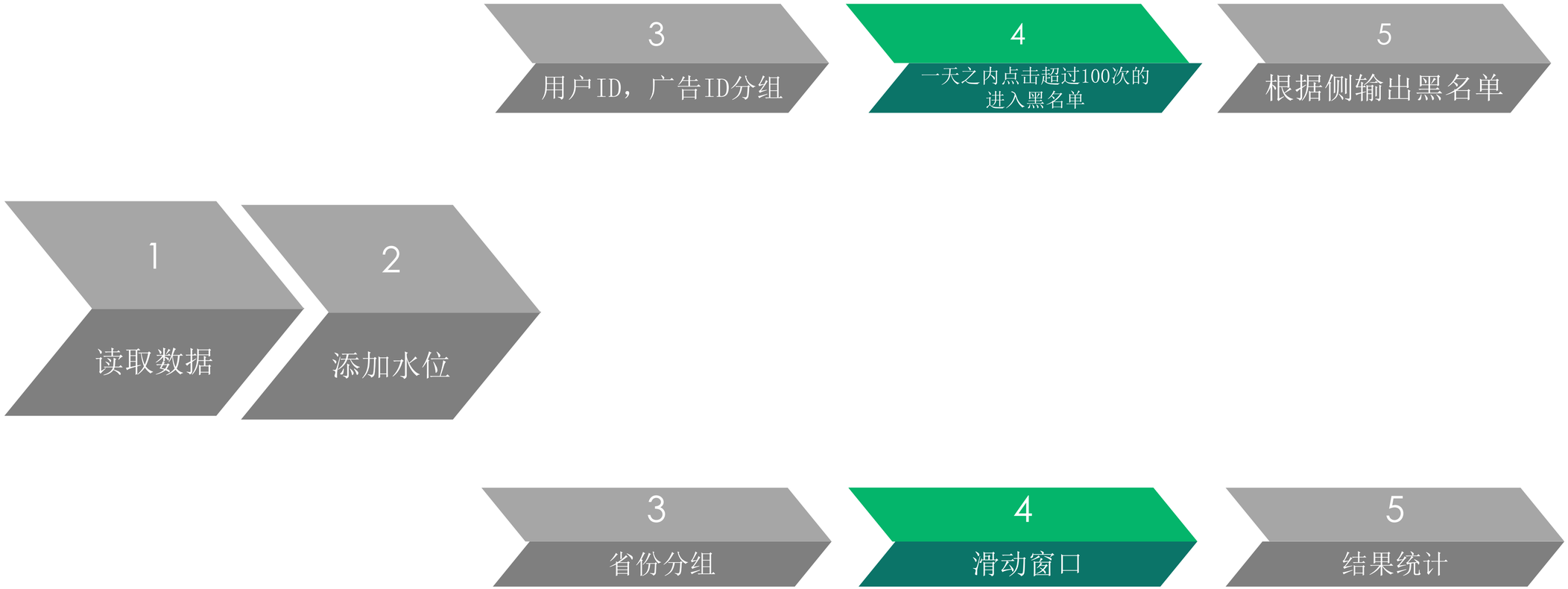
01 实时统计广告点击

需求分析

1. 实时生成黑名单(同一个用户，同一个广告点击)
2. 每隔5秒统计最近1小时的各省份的广告点击

数据展示

用户ID	广告ID	省份	城市	时间
543462	1715	广东	深圳	1511658600


1
读取数据
2
添加水位
3
用户ID，广告ID分组
4
一天之内点击超过100次的
进入黑名单
5
根据侧输出黑名单
3
省份分组
4
滑动窗口
5
结果统计

实现思路

03 效果展示

```
BlackListWarning(937166,1715,Click over100 times)
```

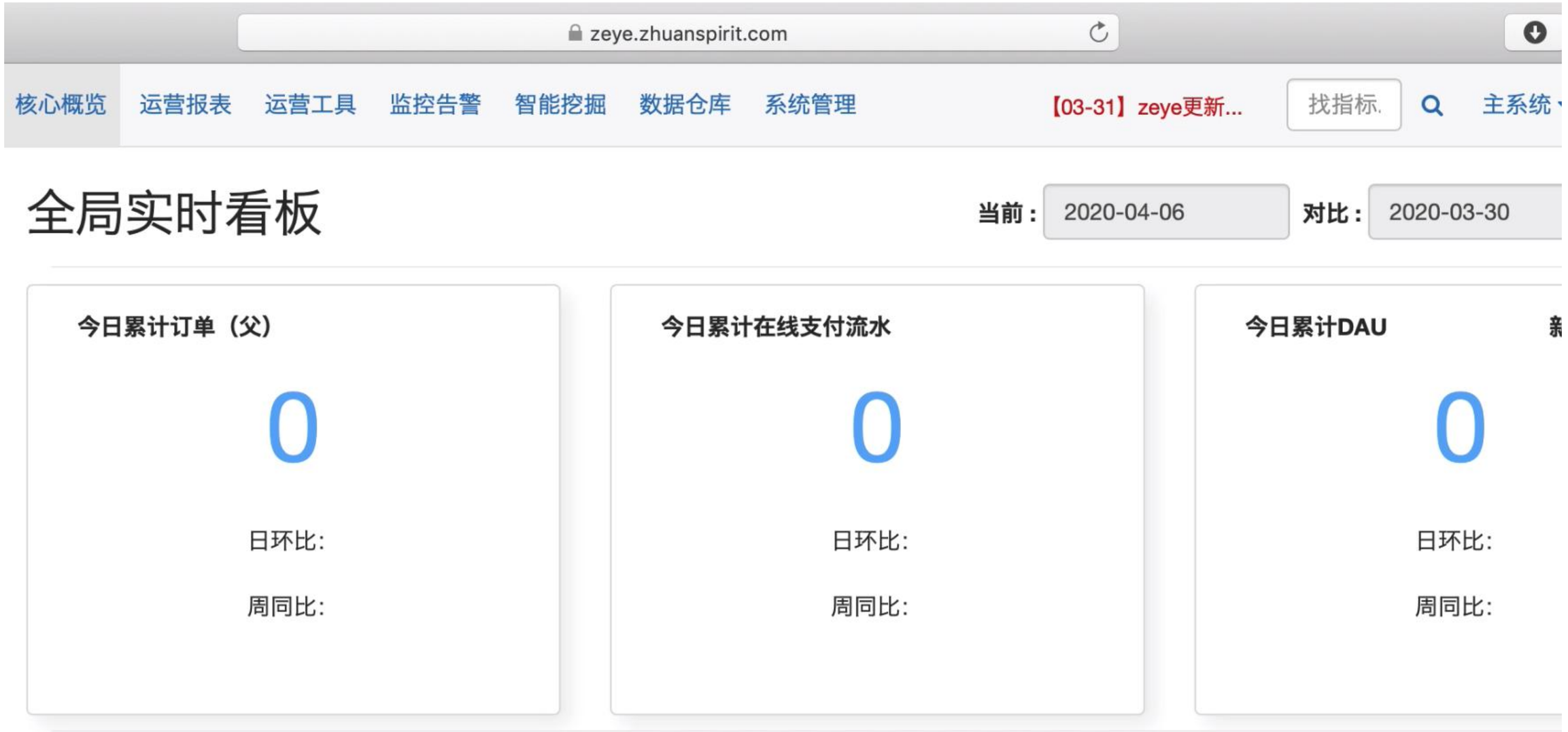
```
CountByProvince(2017-11-26 09:06:05.0,上海,1)
CountByProvince(2017-11-26 09:06:05.0,北京,2)
CountByProvince(2017-11-26 09:06:05.0,广东,4)
CountByProvince(2017-11-26 09:06:10.0,上海,1)
CountByProvince(2017-11-26 09:06:10.0,北京,2)
CountByProvince(2017-11-26 09:06:10.0,广东,4)
CountByProvince(2017-11-26 09:06:15.0,上海,1)
CountByProvince(2017-11-26 09:06:15.0,北京,2)
CountByProvince(2017-11-26 09:06:15.0,广东,4)
```

03 思考

整体概况	优品DAU	有效发布商品数	支付订单量	支付GMV（元）	收入（元）	手机业务支付订单量	手机业务支付GMV（元）	3C业务支付订单量	3C业务支付支付GMV（元）	pop店付支付订单量	pop店付支付GMV（元）

05 需求分析

01 zeye实时指标统计



06 订单统计/GMV统计

需求分析

实时统计每日订单

数据展示

订单编号	订单金额	订单状态	用户id	支付方式	支付流水号	创建时间	操作时间
6624079083	170	6	45869074	OTHERS	6240524417	2020-09-18 00:00:30	2020-09-18 01:52:37

实现思路

今日累计订单（父）

0

日环比:

周同比:

今日累计在线支付流水

0

日环比:

周同比:

1.时间纪元

所谓的"时间纪元"就是1970年1月1日0时0分0秒，指的是开始的时间。比如Java类代码：

Date date = new Date(0);

System.out.println(date);

打印出来的结果：

Thu Jan 01 08:00:00 CST 1970

也是1970年1月1日，实际上时分秒是0点0分0秒，这里打印出来的时间是8点而非0点，原因是存在系统时间和本地时间的问题，其实系统时间依然是0点，只不过我们的电脑时区设置为东8区，故打印的结果是8点。

只需要将时区设置为GMT+0，即可打印出0点0分0秒

System.setProperty("user.timezone","GMT+0");

实际上时区问题都是在此时间纪元基础上加/减一定的offset。

2.Flink时间

说java纪元跟本文将的flink时间问题有啥关系呢?

Flink在使用时间的这个概念的时候就是基于时间纪元这个概念的。比如首先，我们的时区是东八区，在我们的视野中UTC-0时间应该加8小时的offset，才是我们看到的时间，所以在使用flink的窗口的时候往往比我们当前的时间少8小时。

由于Flink默认窗口时区是UTC-0，其他地区需要指定时间偏移量调整时区，在Flink某些低版本中(例如1.6.2等)，官方文档用负时间偏移量

TumblingEventTimeWindows.of(Time.days(1),Time.hours(-8))
会出现以上报错

解决方法：

使用TumblingEventTimeWindows.of(Time.days(1),Time.hours(16))代替

注：在此 https://github.com/apache/flink/pull/5376 已经修复。最新的版本不会报错。

需求分析

统计每日的PV

统计各个品类的PV

统计每日各个品类的订单数

统计各个品类的每日订单金额

统计每日的GMV

统计各个品类每日GMV

.....

07 UV/DAU统计

01 实时统计DAU

需求分析

每小时UV统计

数据展示

用户编号	商品编号	品类编号	用户行为	访问时间
4675130	12437439	4296588	P	1601688552

收藏 | 146 | 82

DAU 编辑

DAU(Daily Active User)日活跃用户数量。常用于反映网站、互联网应用或网络游戏的运营情况。DAU通常统计一日（统计日）之内，登录或使用了某个产品的用户数（去除重复登录的用户），这与流量统计工具里的访客（UV）概念相似。

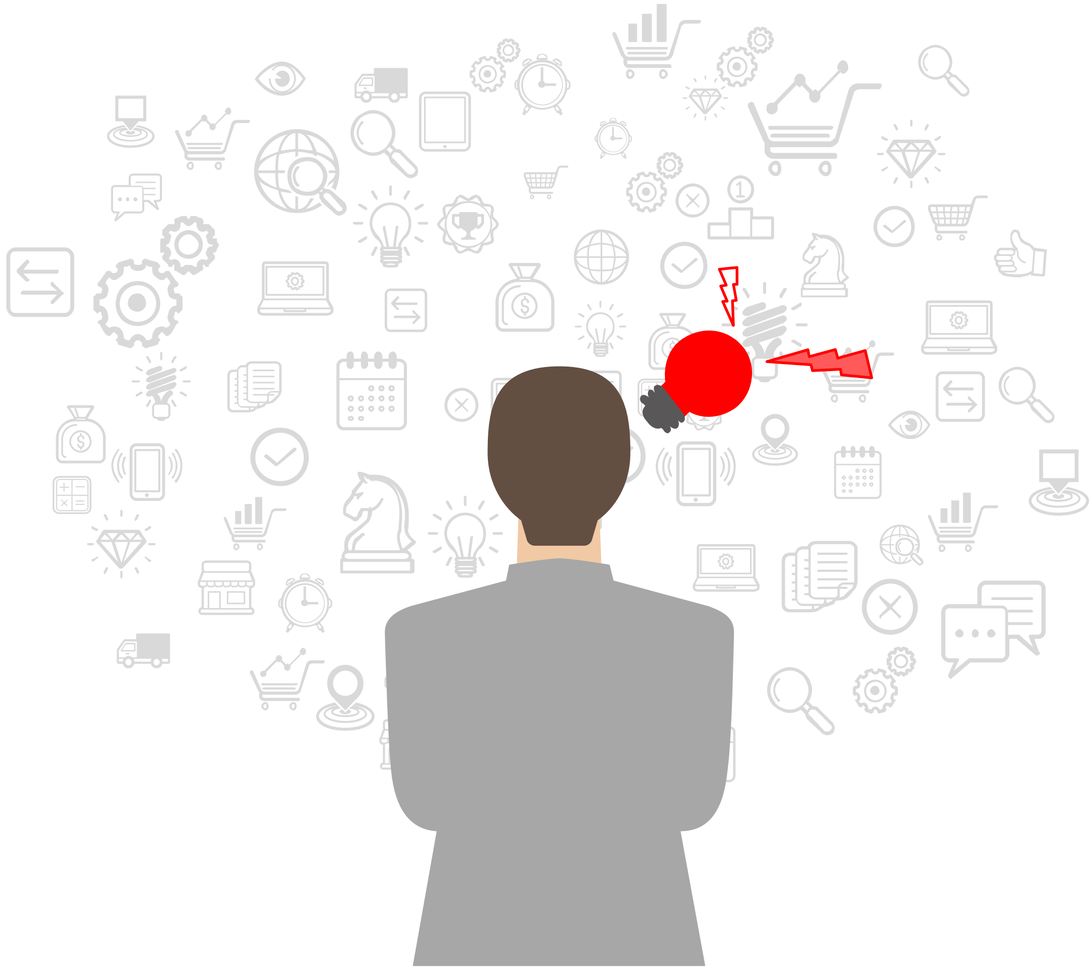
实现思路

01 效果展示

```
UvInfo(2020-10-03 10:00:00.0,17416)
UvInfo(2020-10-03 11:00:00.0,13)
```

01 问题

数据量大了，内存不够?



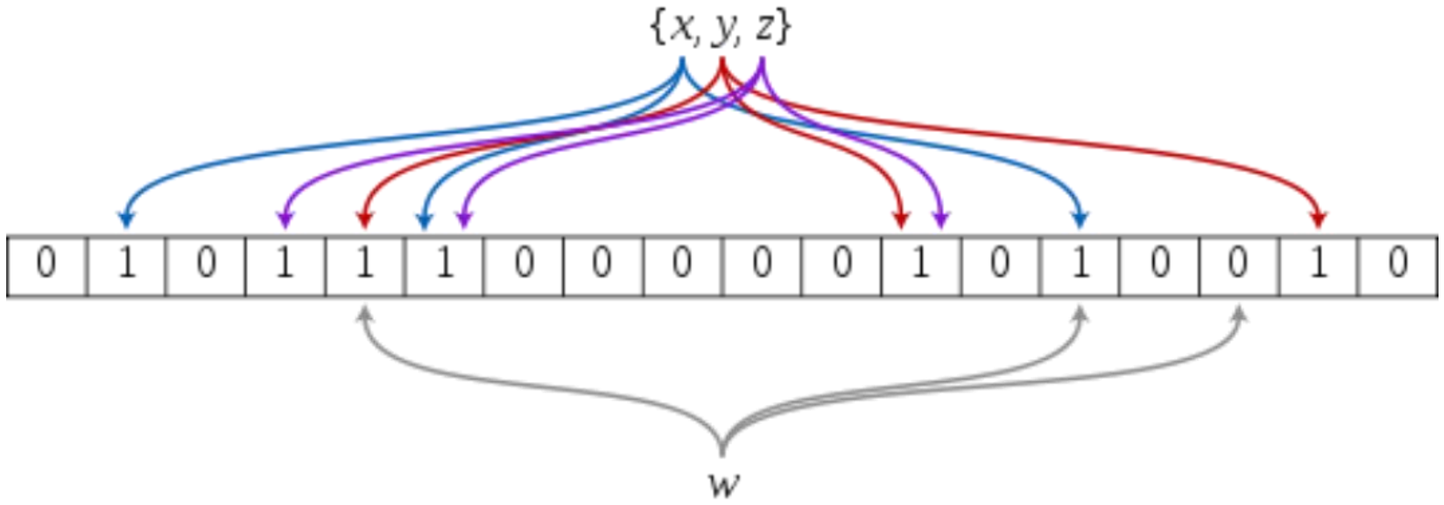
01 引入布隆过滤器

直观的说，bloom算法类似一个hash set，用来判断某个元素（key）是否在某个集合中。

和一般的hash set不同的是，这个算法无需存储key的值，对于每个key，只需要k个比特位，每个存储一个标志，用来判断key是否在集合中。

算法：

1. 首先需要k个hash函数，每个函数可以把key散列成为1个整数
2. 初始化时，需要一个长度为n比特的数组，每个比特位初始化为0
3. 某个key加入集合时，用k个hash函数计算出k个散列值，并把数组中对应的比特位置为1
4. 判断某个key是否在集合时，用k个hash函数计算出k个散列值，并查询数组中对应的比特位，如果所有的比特位都是1，认为在集合中。

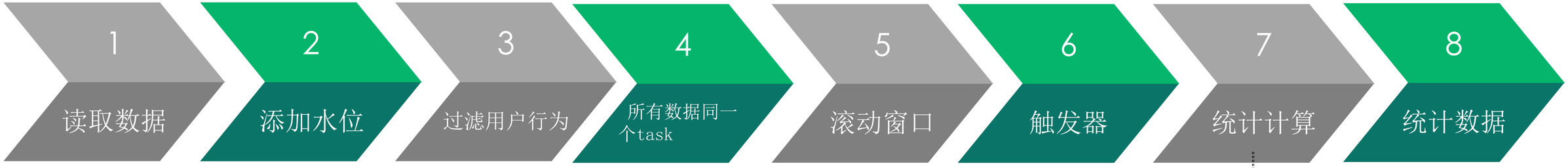


01 引入redis

在我们平时开发过程中，会有一些 bool 型数据需要存取，比如用户一年的签到记录，签了是 1，没签是 0，要记录 365 天。如果使用普通的 key/value，每个用户要记录 365个，当用户上亿的时候，需要的存储空间是惊人的。为了解决这个问题，Redis 提供了位图数据结构，这样每天的签到记录只占据一个位，**365 天就是 365 个位**，46 个字节 (一个字节有8位) 就可以完全容纳下，**这就大大节约了存储空间**。

位图不是特殊的数据结构，它的内容其实就是普通的字符串，也就是 byte 数组。我们可以使用普通的 get/set 直接获取和设置整个位图的内容，也可以使用位图操作 getbit/setbit等将 byte 数组看成「位数组」来处理。

02 使用布隆过滤器实时统计UV



实现思路

1. 写一个布隆过滤器
2. 计算当前用户编号hash值
3. 计算在布隆过滤器的位置
4. 根据上个步骤更新结果

03 效果展示

```
(2020-10-03 10:00:00.0,17416)
(2020-10-03 11:00:00.0,13)
```

奈学教育，一个有干货更有温度的教育品牌

出品： 奈学教育